Zaawansowane Metody Uczenia Maszynowego
projekt zespołowy

sem.zimowy 2020/2021

Dr inż. Grzegorz Sarwas

Celem projektu jest opracowanie modelu regresji pokazującego trend w wybranym przez studentów zbiorze danych wraz z jego analizą. Do zdobycia jest **35 pkt**.

1. **Wartość merytoryczna – 30 pkt.**
   Oceniana na podstawie raportu z przeprowadzonych prac – termin do końca 1 tygodnia sesji.
2. **Prezentacja – 5 pkt.**
   Analiza eksploracyjna posiadanego zbioru/wycinka zbioru danych i postawienie tezy/zadania badawczego mającego na celu opracowanie modelu regresji dla opisanych danych.

Zasady realizacji projektu i sposób oceniania:

1. Projekt realizujemy w zespołach 3 osobowych. Każdy zespół ma za zadanie poszukać ogólnodostępnego zbioru danych:
   a. Dominic's Dataset: https://www.chicagobooth.edu/research/kilts/datasets/dominicks
   b. Dane GUSu w Polsce, jak i za granicą: https://stat.gov.pl/podstawowe-dane/
   c. Dane Eurostatu
   d. Dane giełdowe, medyczne, astronomiczne itp.

   Można także poszukać dowolnego innego, sensownego zbioru danych w Kaggle Datasets, jak również wśród różnych danych prezentowanych przez firmy rządowe lub pożytku publicznego. Zależy nam na surowych danych bez postawionego problemu, po to by postawienie jakiegoś zagadnienia wynikało z przeprowadzonej analizy danych.
2. Pierwszą częścią projektu jest eksploracyjna analiza danych. Jej celem jest sprawdzenie zależności między posiadanymi danymi, zbadanie ich zakresów i stopnia zmienności, analiza stopnia wypełnienia danych (5 pkt.). Analiza danych musi być opatrzona należytą wizualizacją (5 pkt.). Wszystkie analizy i wykresy muszą być opisane i podsumowane. **Wynikiem przeprowadzonych analiz ma być postawienie hipotezy badawczej mającej na celu znalezienie relacji między zmiennymi objaśniającymi, a zmienną objaśnianą.** (5 pkt.)
3. Druga część projektu związana jest z opracowaniem modelu regresji wynikającego z postawionej hipotezy. Na początku należy dokonać imputację brakujących danych (jeśli jest wymagana), dokonać opracowania nowych cech i przygotować dane do dalszych prac (5 pkt.). Dodatkowo należy dokonać doboru cech 4 różnymi poznanymi na wykładzie i ćwiczeniach laboratoryjnych metodami i przeanalizować otrzymane wyniki (5 pkt.). Należy porównać modele otrzymane różnymi metodami i dla najlepszego modelu należy zastosować metody regularyzacji w celu ograniczenia jego wariancji. Należy wyciągnąć wnioski na podstawie otrzymanych wyników/zależności (5 pkt.).
4. Przygotować 10 min. prezentację z przeprowadzonych prac projektowych, która podlegać będzie publicznej obronie w pierwszym tygodniu sesji – termin zostanie ustalony, jak będzie już znany plan sesji (5 pkt).

Warunkiem zaliczenia jest przygotowanie raportu i przedstawienie prezentacji tj. bez przedstawienia tych dwóch elementów zaliczenie nie będzie możliwe. Podane punkty w obu tych przypadkach można zdobyć za staranność wykonania, jasność i czytelność wypowiedzi (pisemnej/ustnej), sposób prezentacji problemu i rozwiązania, sensowność sformułowanych na zakończenie wniosków i innych czynników branych zwykle pod uwagę przy ocenianiu tego typu aktywności.